CIS: Quick Guide for Users

Dataset:	Community Innovation Survey (CIS)
Dates available:	1994-1996 (CIS2), 1998-2000 (CIS3), 2002-2004 (CIS4), 2005-2007 (CIS 2007 mini)
Survey questions:	Innovation
Collected by:	CIS2 – CIS4, DTI; CIS2007, DIUS
Link fields:	IDBR reference
Legal restrictions:	Voluntary survey: not covered by STA, covered by NS Code of Practice.

Quick summary

The Community Innovation Survey (CIS) is a voluntary postal survey carried out by ONS on behalf of the DTI (now DIUS). Eurostat proposes an initial questionnaire and the DIUS adds questions. ONS randomly selects a stratified sample of firms with more than 10 employees, drawn from the Inter-Departmental Business Register (IDBR) by SIC92 2-digit class and 8 employment size bands. The survey covers both the production (manufacturing, mining, electricity, gas and water, construction) and the service sectors. The retail sector has been excluded from the survey as this sector has been a poor responder in previous surveys and, generally, has shown very little innovation.

An enterprise is defined as being innovation active within the period 1998-2000 if it has:

- Introduced a new or significantly improved good, service or process;
- Engaged in innovation projects that are not yet complete;
- Engaged in longer term innovation activities such as basic R&D or technology watch;
- Had innovation related expenditure
- Formally co-operated with other enterprises or institutions on innovation.

Sampling frame

The Second Community Innovation Survey (CIS 2) covered the period 1996-98. There is little information available about this survey now.

CIS 3 was in the field twice. The first wave sampled 13,340 enterprises, the second top-up covered 6,285 to make the sample representative at the regional level. The CIS 3 covers the period 1998-2000. Of the total 19,625 enterprises to which the survey was sent, 8,172 responded (Table 1, row 1), achieving a response rate of 42%.

The 2005 survey (CIS 4) is the largest innovation survey so far conducted, sent to 28,000 UK enterprises and achieving a 58% response rate.

The latest CIS 2007 survey contains approximately 14,000 observations. Note this is a mini ‘in between’ survey. The next full survey of the CIS is the CIS 5.

CIS2007 received a higher response rate than any of the previous surveys. Importantly, approximately half of the respondents to this latest survey **had also previously responded to the CIS4**. DIUS have kindly provided the VML with a CIS4/CIS2007 panel data set. This file contains the observations for approximately 7000 respondents who completed both surveys.

Organisation of files

The data is received from the DTI, substantially cleaned, and made available as the files *cis2_clean_01.dta*, *cis3_clean_01.dta* and *cis4.dta*. These are held within the clean drive (U:\CIS). For details of cleaning, see the audit documents and other files held within the documentation drive (X:\).

The CIS 3 had previously been linked to the 2000 ARD by reporting unit reference, creating the dataset *cis3_ard_2000.dta*. It is held within the *CIS_data archive.zip* within the source drive. This linkage includes a number of assumptions about behaviour, and may not be appropriate for all users. While users may use this file, it may be worthwhile to create your own linked panel to the latest version of ARD2. The default linking mechanism is simply to use the Reporting Unit reference **dlink_ref2** common to both files (and most other BDL files).

Known data issues

There are a large number of concerns about the data, primarily as the result of obvious inconsistencies. As a result, a lot of marker variables were created during the cleaning process. These are summarised in the *variable reference.xls* held within the documentation drive (X:\CIS) for more details please refer to the audit document.

The dataset has not been thoroughly investigated yet, nor have the inconsistencies been corrected. Some of these are probably due to the design of the survey, some to misunderstandings by the respondents. Some are probably clerical errors (eg 0.6% of innovation expenditure breakdowns in the CIS3 do not add up to the reported total).

A specific issue is that the CIS2 contains three duplicated RU references. For each of the duplicated references, there is no clear reason to pick one or other of the duplicates. They are therefore left in the dataset but should be removed or reduced before merging on RU references.

Overall, these surveys require users to use them cautiously. A detailed study of the survey forms and documentation is recommended before starting analysis.

UK Innovation Survey 2009
User Guide

Contents

Introduction

The UK Innovation Survey (UKIS) provides the main source of information on business innovation in the UK. The survey data is a major resource for research into the nature and functioning of the innovation system and for policy formation. It is used widely across government, regions and by the research community.

The survey is funded and developed by the Department of Business, Innovation and Skills (BIS) and administered by the Office for National Statistics (ONS) with assistance from the Northern Ireland Department of Enterprise, Trade and Investment (DETI).

The UK Innovation Survey also represents the UK's contribution to the Europe-wide Community Innovation Survey (CIS).

This user guide is based on the sixth iteration of the survey: UKIS 2009 (sometimes referred to as CIS6 or CIS 2008.)

Overview

The purpose of the survey is to collect information about businesses innovation in the UK.

Like many innovation surveys across Europe, the UK innovation Survey follows general guidelines set out in an OECD publication known as the Oslo manual (OECD 2005)[1]. This manual provides guidelines on the conduct of innovation surveys, including statistical procedures and a review of the range of concepts that fall together under the umbrella term "innovation".

The survey is based on a core questionnaire developed by Eurostat[2] and Member States and covers a broad range of policy interests including:

- General business information
- Innovation activity
- Goods, services and process innovation
- Context for innovation, and
- General economic information.

The UK survey was originally conducted every four years, but since 2005 has been conducted biennially. Previous Community Innovation surveys took place in 2007, 2005, 2001.

In the UK, the survey is voluntary and collected by means of a postal questionnaire.

[1] http://www.oecd.org/document/33/0,3343,en_2649_34273_35595607_1_1_1_37417,00.html
[2] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Community_innovation_survey

Coverage and Sampling

The UK innovation survey consists of a nationally representative sample of businesses with 10 or more employees in sections C-K of the Standard Industrial Classification (CIS) 2003.

The sample is a stratified design drawn from the Inter-Departmental Business Register (IDBR) with Neyman allocation used to determine the sample size in each stratum. Overall, roughly ten per cent of the target population in sampled.

Stratification was based on three variables:

1. **Region -** All regions and countries in the UK (9 Government Office Regions in England plus Scotland, Wales and Northern Ireland) are covered

    - North East
    - North West
    - Yorkshire and The Humber
    - East Midlands
    - West Midlands
    - East of England
    - London
    - South East
    - South West
    - Wales
    - Scotland
    - Northern Ireland

2. **Division -** Coverage of the following sectors in the target population:

    - Div 10-14 - Mining and quarrying [a]
    - Div 15-22 - Manufacture of food, clothing, wood, paper, publishing and printing [a]
    - Div 23-29 - Manufacture of fuels, chemicals, plastics metals & minerals [a]
    - Div 30-33 - Manufacture of electrical and optical equipment [a]
    - Div 34-35 - Manufacture of transport equipment [a]
    - Div 36-37 - Manufacture not elsewhere classified [a]
    - Div 40-41 - Electricity, gas and water supply [a]
    - Div 45 - Construction [b]
    - Div 50 - Sale, maintenance and repair of motor vehicles [b]
    - Div 51 - Wholesale trade [a]
    - Div 52 - Retail trade (exc. cars & bikes) and repair [c]
    - Div 55 - Hotels & restaurants [c]
    - Div 60-63 - Transport & storage [a]
    - Div 64.1 - Post & courier activities [a]
    - Div 64.2 - Telecommunications [a]
    - Div 65-67 - Financial intermediation [a]
    - Div 70 - Real estate [b]
    - Div 71 - Renting [b]
    - Div 72 - Computer & related activities [a]
    - Div 73.1 - R&D (natural sciences & engineering) [b]
    - Div 73.2 - R&D (social sciences & humanities) [b]
    - Div 74.2 - Architectural & engineering activities [a]
    - Div 74.3 - Technical testing and analysis [a]
    - rest of 74 - Other business activities (exc. SIC 74.2 & 74.3) [c]
    - SIC 92.11 - Motion picture and video production [d]

Note:
[a] Denotes required under an EU regulation on innovation statistics.
[b] Denotes additional sectors covered in the 2001 UK survey onwards
[c] Denotes additional sectors covered in the 2005 UK survey onwards
[d] Denotes additional sector covered in the 2007 UK survey onwards

3. **Business size -** All enterprises with 10 or more employees are included in the target population:

    - Small 10-49 employees
    - Medium 50-249 employees
    - Large 250 or more employees

Additionally, to ensure representativeness, the following conditions were also introduced:

- A census for all large firms (250+ employees) is taken,
- A census of SMEs in SIC 40-41 and 73.2, where the population is particularly small, is taken, and
- A cap on the number sampled from SIC 50-51, 52, 55 and "rest of 74", where the population is particularly large, is taken.

The questionnaire

The questionnaire content is determined by Eurostat regulatory requirements and BIS via the UKIS project board. Eurostat provide the core (harmonised) questionnaire to ensure European Union data requirements are met and provide the basis for comparisons with other countries. BIS are responsible for identifying the need for new questions or changes to existing questions so that the UKIS continues to provide a means to measuring the level, types and trends in innovation activity in the UK and provide the empirical evidence to support policy.

The core questionnaire covers a broad range of innovation-related concepts, including:

- Details of any innovation-related activities such as R&D, acquisition of equipment, training, design etc.
- Information on innovations in business strategies and practices
- Product innovation
- Process innovation
- Abandoned and incomplete innovation activities
- The context for innovation e.g. increase range of goods or services, entering new markets
- Cooperation agreements
- The factors constraining innovation.

Some core questions are only required every fours year, rather than every two – these are indicated in the Variable details.

Each business in the sample is sent the questionnaire and a leaflet containing some results from the previous survey. The survey is voluntary; however, businesses receive two postal reminders and can be contacted by telephone to complete the questionnaire or to validate responses.

Following cogitative testing, the 2009 survey was sent out to businesses at the end of March 2009 and remained active until September that same year.

Response and weighting

Valid responses were received from 14,281 enterprises to give a response rate of around 50 per cent. Accordingly, weighting is used to compensate for the businesses that did not respond to the survey and those not selected for the sample.

Weighting allocates a "weight" to each business, ensuring that the respondents are representative of the target population as a whole, in terms of region, division and business size.

Two weights are available to users:

1. **Business weights** – these are frequency weights that indicate the number of enterprises a respondent represents within their strata. On average, each respondent represents 13 enterprises in the population.
2. **Employment weights** – these are frequency weights that indicate the number of enterprises a respondent represents according the number of employees in their business and the total number employees within their strata (taken from the IDBR.)

Variable coding

Details of the coding used for each variable are set out below under the Details of variables. Typically, variable labels adhere to the convention: 1 – yes, 0 – no; and 3 – High, 2 – Medium, 1 – Low. Unless otherwise specified in this guide, the coding for "not answered" and "not applicable" are -8 and -9, respectively.

Note on missing values

The routing used throughout the questionnaire means that businesses are not required to answer every question – some questions are not applicable.

In these instances (where a question is not applicable), responses for that particular business are coded as -9 and defined as "missing" – i.e. their values will not be included in any analysis.

There are, however, occasions where a question is appropriate to the business but a certain "category" or "response" contained in the question may not be. In this instance, their response is considered as "valid".

Like with many surveys, there are occasions when a respondent does not or cannot answer a particular question and leaves it blank. This type of response is coded as -8 and labelled "not answered". The numbers of businesses that left a particular question blank, on some occasions, is considerable and for this reason "not answered" is not treated as an invalid (or "missing") response.

Innovation Concepts

Innovation, for the purpose of this survey, is defined as new or significantly improved goods or services and/or the processes used to produce or supply these.

Product innovation – bringing to the market or into use by business, new and improved products, including both tangible goods and the provision of services. The degree of innovativeness is shown by the distinction between products new just to the business or which are also new to the market.

Process innovation – significant changes in the way that goods or services are produced or provided, again differentiating between processes new to the business only or also new to the industry.

Innovation-related activities – categories of innovation directed investment such as: R&D, capital goods and software acquisition, design activity, for implementing current innovations or directed to future product or process changes

New to market – the introduction of a new good or service to the market before competitors.

New to this business – introduction of a new good or service that was essentially the same as a good or service already available from competitors.

Wider or "Strategic" innovation – new and significantly improved forms of organisation, business structures or practices aimed at improving internal efficiency or effectiveness of approaching markets and customers.

Question/Variable look-up table

Question No.	Variable name[3]	Short description
1	@210-@240	Geographic market
2	@410-@440	Business changes
3	@1-@4	Business objectives
4	@1310-@1374	Innovation-related activity
5	@1410-@1470	Innovation-related activity expenditure in 2008
6	@2310-@2340	Business strategy and practices
7	@510-@520	New or significantly improved products
8	@601-@630	Who developed products
9	@710-@720	New to market/new to business products
10	@810-@840	New to market/new to business products 2008 turnover
11	@900	New or significantly improved processes
12	@1010-@1030	Who developed processes
13	@1100	New to industry processes
14	@1510-@1520	Incomplete or abandoned innovation activities
15	@1501	Innovation marker
16	@1210-@1280	Innovation factors
17	@1601-@1690	Importance of information
18	@1811-@1874	Innovation cooperation
20	@2011-@2030	Reasons for not innovating
21	@1901-@1911	Innovation constraints
22	@2130-@2150	Innovation protection
23	@2210-@2240	Financial support
24	@2410-@2420	Business turnover
25	@2510-@2520	Average employees
26	@2610-@2620	Skills

[3] The @ prefix is used for variables in the SPSS dataset. For Stata, variable names are prefixed with __.

Sample Design in the UK CIS4

Introduction

The Community Innovation Survey (CIS) is traditionally based on a stratified random sample drawn from the ONS Inter-Departmental Business Register (IDBR). A considered sample design is essential to ensure that the data collected are as precise as possible and representative of the population of interest. Key users of disaggregated data include regional analysts in the Regional Development Agencies (RDAs) and the Devolved Administrations (DAs), as well as analysts of industrial statistics. This paper considers appropriate methodologies, goes on to investigate the structure of the population counts in the IDBR, and finally makes a proposal for a design for the sample.

The CIS4 population

CIS4 will again be based on a stratified random sample using the same stratification variables as in CIS3, namely sector, region and sizeband.

Sector coverage

Coverage of the following sectors in the target population is required under an EU regulation on innovation statistics:

- SIC 10-14 - Mining and quarrying
- SIC 15-37 - Manufacturing
- SIC 40-41 - Electricity, gas and water supply
- SIC 51 - Wholesale trade
- SIC 60-64 - Transport, storage and communication
- SIC 65-67 - Financial intermediation
- SIC 72 - Computer and related activities
- SIC 74.2 - Architectural and engineering activities
- SIC 74.3 - Technical testing and analysis

All of the above sectors were also covered in CIS3 in the UK.

In addition to these, it was agreed at the CIS4 project board meeting of 2 July 2004 that the following additional sectors would be covered in the UK survey:

- SIC 45 - Construction*
- SIC 50 - Sale, maintenance and repair of motor vehicles*
- SIC 52 - Retail trade
- SIC 55 - Hotels and restaurants
- SIC 70 - Real estate*
- SIC 71 - Rental of machinery and equipment*
- SIC 73 - R&D*
- Remainder of SIC 74 (i.e. excl. 74.2 and 74.3) – Other Business Activities

(* denotes sectors that were covered in CIS3 in the UK)

The detailed stratification is considered towards the end of this paper.

Region/country coverage

As in CIS3, all regions and countries in the UK (9 Government Office Regions in England plus Scotland, Wales and Northern Ireland) will be covered and these 12 areas used in the regional dimension of the stratification.

Sizebands

All enterprises with 10 or more employees will be included in the target population and the following strata used: 10-49, 50-249, 250-499, 500-999 and 1000+.

Statistical units

The sample will be drawn at the level of the reporting unit, in line with other major, relevant business surveys (e.g. BERD, ABI). Some users have noted that by sampling at the reporting unit level we often lose sight of diffusion of innovation within the company (at the local unit level), and that this might lead to cross-border innovation not being picked up. Therefore a file detailing all local units within the selected reporting units will be extracted from the IDBR at the time of the main sample extraction, and this will accompany the main, reporting unit-level, dataset. Note that it may also be possible to obtain information about the enterprise groups for the selected reporting units, and all reporting units within them.

Sample selection methodology

At the CIS4 project board meeting of 2 July 2004 it was agreed that the sample would be selected using a Neyman allocation based on firms' innovation active rates in CIS3. This will help ensure that optimal precision will be obtained in CIS4, assuming that the observed patterns are broadly similar over time.

Sample size

The ONS has agreed to a size limit of 28,000 firms for the UK sample. The working assumption is that the final sample will be very close to 28,000.

Population counts

In order to consider an appropriate sample design, it is first useful to look at the population counts in a recent IDBR extract (taken on 8 December):

The CIS4 population is heavily skewed towards the smaller firms, with roughly two thirds of firms based in the service sector. Of note, roughly 40% of all firms in the identified population are small firms (10-49 employees) in four service sectors, namely SIC 50-51 & 52 (wholesale and retail), 55 (hotels and restaurants) and "rest of 74" (other business activities).

There are also considerable differences in the sizes of the regions, with the likes of London and the South East being nearly five times larger than regions such as Northern Ireland and the North East:

Table 2: Number of firms in UK CIS4 population by region

	Employees size-bands						
Region	**10-49**	**50-249**	**250-499**	**500-999**	**1000+**	**250+**	**Total**
North East	4,525	885	125	65	35	225	5,635
North West	15,120	2,905	390	160	155	705	18,735
Yorkshire and The Humber	11,445	2,235	270	150	110	530	14,205
East Midlands	10,195	1,960	235	120	90	445	12,600
West Midlands	12,585	2,400	295	160	125	580	15,560
East	12,995	2,295	280	130	105	515	15,805
London	19,915	3,935	535	330	325	1,190	25,035
South East	19,660	3,565	465	225	275	965	24,195
South West	11,850	2,020	240	110	95	445	14,315
Wales	4,890	1,020	140	70	30	240	6,145
Scotland	10,070	1,965	255	135	95	485	12,520
Northern Ireland	4,930	880	100	45	30	175	5,985
UK total	**138,190**	**26,060**	**3,325**	**1,690**	**1,470**	**6,485**	**170,735**

Response rates

In CIS3, response rates varied according to the size of the firm:

Table 3: CIS3 UK response rates

	Numbers of businesses in sample			Number of businesses with response			Response rates		
Wide SIC group	**SME**	**Large**	**Total**	**SME**	**Large**	**Total**	**SME**	**Large**	**Total**
10-14	275	49	324	111	16	127	40%	33%	39%
15-22	1,641	702	2,343	768	237	1,005	47%	34%	43%
23-29	1,959	752	2,711	904	217	1,121	46%	29%	41%
30-33	954	340	1,294	422	105	527	44%	31%	41%
34-35	735	204	939	285	59	344	39%	29%	37%
36-37	941	113	1,054	388	55	443	41%	49%	42%
40-41	69	46	115	33	20	53	48%	43%	46%
45	1,929	271	2,200	829	118	947	43%	44%	43%
51	2,008	383	2,391	918	123	1,041	46%	32%	44%
60-64	1,327	446	1,773	601	172	773	45%	39%	44%
65-67	823	312	1,135	331	74	405	40%	24%	36%
70-74	2,751	572	3,323	1,194	192	1,386	43%	34%	42%
Total	**15,412**	**4,190**	**19,602**	**6,784**	**1,388**	**8,172**	**44%**	**33%**	**42%**

Source: CIS3

Response rates were reasonably consistent across industrial sectors, ranging from 36% to 46%. However, in nearly all sectors SMEs were more likely to respond than large firms.

A note on weighting

Sample-based business survey data are commonly grossed up to make inferences about the whole business population, and in order to do this a set of weights is applied to the raw data. There are two potential sets of weights that can be used, namely business weights and employment weights, to adjust estimates to account for firms not included in the sample. In the past, CIS data have typically been

weighted using business weights, which are calculated as the inverse sampling fraction of the number of firms in each cell of the stratification. Using business weights, each firm in the population carries an equal weight. This can provide basic measures of population performance, but does not correct for differing firm sizes. Using employment weights, which are calculated using the inverse sampling fraction of employees in each cell of the stratification, allows data to be grossed up and corrected for firm size. So the greater the size of the firm, the larger the weight it carries in the employment weighted results. Employment weighted data provide a more robust means of estimating whole economy performance.

Given large firms' dominance of economic activity in the UK (the 6,500 large firms in the CIS4 population account for over 60% of total employment and around 70% of turnover in the population) and the arguments for using employment weights to provide better measures of whole-economy performance, a census of large firms is recommended. For the SME strata (10-49 and 50-250), sampling should be carried out using the Neyman allocation.

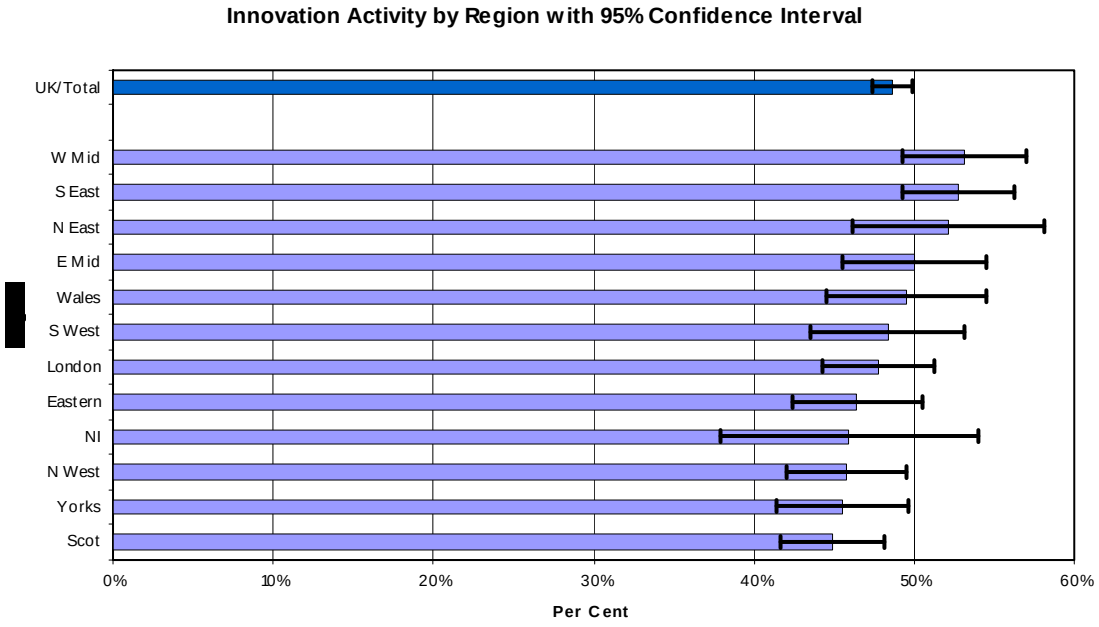
Recommendation: Take a census for large firms (250+ employees) and sample based on optimal allocation for SMEs

Stratification and issues of sample design

Regional stratification

A DTI paper – "Assessing the accuracy of published results of the UK CIS3" – revealed that there are large disparities in the precision of the published CIS3 results for the 12 regions (see chart 1).

Chart 1



Given that the results of CIS4 are likely to be used for regional benchmarking, one conclusion of the paper was that a greater level of regional balance in the sample would be desirable to allow for more precise regional comparisons. A slightly larger sample is necessary for the larger regions to achieve a given level of precision.

It can be shown that the following regional distribution of the sample would be necessary to ensure 95% confidence intervals no wider than ±3%:

Table 4: Required regional distribution to achieve ±3% precision of innovation active rate estimation in CIS4

Region	N (CIS4)	p (CIS3)	q (CIS3)	Required achieved sample size, n (for +/- 3% precision)	Response rate (CIS3)	Required total sample size	Sample size in CIS3
North East	5,635	0.52	0.48	896	40%	**2,230**	1,105
North West	18,735	0.458	0.542	1,003	42%	**2,416**	2,026
Yorks and Humber	14,205	0.455	0.545	985	42%	**2,320**	1,773
East Midlands	12,600	0.5	0.5	984	43%	**2,310**	1,641
West Midlands	15,560	0.531	0.469	995	43%	**2,306**	1,696
East	15,805	0.465	0.535	995	43%	**2,337**	1,761
London	25,035	0.478	0.522	1,022	38%	**2,693**	2,567
South East	24,195	0.527	0.473	1,019	41%	**2,506**	2,488
South West	14,315	0.483	0.517	992	40%	**2,503**	1,567
Wales	6,145	0.495	0.505	909	45%	**2,015**	840
Scotland	12,520	0.448	0.552	974	46%	**2,097**	1,734
Northern Ireland	5,985	0.459	0.541	901	40%	**2,246**	404
UK	170,735			11,675		**27,978**	19,602

Source: CIS3 and DTI calculations

The calculations in table 4 are based on the following assumptions:

- The proportion of firms in each region that are innovation active is the same as in CIS3;
- Regional response rates in CIS4 are the same as in CIS3

In all regions, an increased sample size would be necessary to achieve the pre-defined level of precision, although in some of the smaller regions the size of the required increase is considerably greater than in the larger regions.

Industrial Stratification

Having proposed a regional stratification, the next consideration is how the sample should be distributed between SIC groups within each region. It has already been proposed that optimal allocation should be used for the SME element of the population. Given that we have shown that there was little variation in response rates across industries in CIS3, it is proposed that the following 23 strata be used in the sample stratification for each region for the SME group:

Table 5: Proposed industrial stratification for CIS4

SIC	Sector Name	Population size	SME	Large
10-14	Mining and quarrying	385	340	45
15-22	Mfr of food, clothing, wood, paper, publishing and printing	12,235	11,520	720
23-29	Mfr of fuels, chemicals, plastics metals & minerals	17,305	16,515	790
30-33	Mfr of electrical and optical equipment	4,105	3,830	275
34-35	Mfr of transport equipment	1,685	1,445	240
36-37	Mfr not elsewhere classified	3,045	2,945	95
40-41	Electricity, gas and water supply	105	75	30
45	Construction	16,800	16,455	350
50-51	Wholesale & commission trade (inc. cars, bikes and fuel)	24,830	24,240	595
52	Retail trade (exc. cars & bikes) and repair	14,085	13,550	540
55	Hotels & restaurants	19,310	18,960	350
60-63	Transport & storage	8,490	8,110	380
64.1	Post & courier activities	580	555	25
64.2	Telecommunications	615	555	55
65-67	Financial intermediation	3,965	3,620	345
70	Real estate	4,745	4,605	140
71	Renting	2,085	2,020	60
72	Computer & related activities	5,290	5,145	145
73.1	R&D (natural sciences & engineering)	540	480	60
73.2	R&D (social sciences & humanities)	70	65	0
74.2	Architectural & engineering activities	4,260	4,150	110
74.3	Technical testing and analysis	350	330	15
rest of 74	Other business activities (exc. SIC 74.2 & 74.3)	25,865	24,735	1,130
Total		**170,735**	**164,250**	**6,485**

Recommendation: Sector stratification as in table 5, with sample selection based on optimal allocation for SMEs

It might also be sensible to take a census of SMEs in certain sectors where the population is particularly small, for example in SIC 40-41 and 73.2, and also to cap the sample in sectors where the population is particularly large, such as SIC 50-51, 52, 55 and "rest of 74".

Sample Design Scenario Analysis

Four scenarios for sample designs are now considered. In each case, it is assumed that the minimum cell size in the sample is 5, and that a census is taken for firms with 250+ employees. The pros and cons of each design are considered below each scenario.

Scenario 1 – "Optimal Regional Allocation"

This scenario considers the sample required to achieve the regional precision levels presented in table 4. Each region is allocated the required sample size using the Neyman allocation.

Table 6 – Sample design under scenario 1

"Optimal Regional Allocation" by region

	10-49	50-249	250+	Total
North East	1684	323	225	**2232**
North West	1302	406	705	**2413**
Yorks and Humber	1366	425	530	**2321**
East Midlands	1491	375	445	**2311**
West Midlands	1261	464	580	**2305**
East	1420	402	515	**2337**
London	1065	438	1190	**2693**
South East	1151	392	965	**2508**
South West	1658	398	445	**2501**
Wales	1271	504	240	**2015**
Scotland	1162	449	485	**2096**
Northern Ireland	1852	221	175	**2248**
UK	**16683**	**4797**	**6495**	**27975**

"Optimal Regional Allocation" by industry

	10-49	50-249	250+	Total
10-14	63	55	45	**163**
15-22	1138	495	720	**2353**
23-29	1495	641	790	**2926**
30-33	489	206	275	**970**
34-35	241	134	240	**615**
36-37	509	130	95	**734**
40-41	40	25	30	**95**
45	1753	321	350	**2424**
50-51	2366	574	595	**3535**
52	1565	202	540	**2307**
55	2136	339	350	**2825**
60-63	998	231	380	**1609**
64.1	73	50	25	**148**
64.2	67	45	55	**167**
65-67	509	126	345	**980**
70	346	98	140	**584**
71	180	61	60	**301**
72	347	126	145	**618**
73.1	60	50	60	**170**
73.2	40	0	0	**40**
74.2	358	88	110	**556**
74.3	64	45	15	**124**
rest of 74	1846	755	1130	**3731**
Total	**16683**	**4797**	**6495**	**27975**

Pros:

- Should provides better levels of regional precision than CIS3, and hence facilitates better regional comparison.
- Gives regional analysts more comprehensive datasets to work with.

Cons:

- Does not provide optimal results from the point of view of UK- and industry-level estimates.
- Larger relative burden on firms in smaller regions, e.g. North East (40% of firms sampled) compared to London (11%).

Scenario 2 – "Optimal National Allocation"

This scenario applies the same methodology as was used in CIS3, namely to apply the Neyman allocation to the whole UK dataset and select the sample to provide optimal national-level estimates.

Table 7 – Sample design under scenario 2

"Optimal National Allocation" by region

	10-49	50-249	250+	Total
North East	807	167	225	**1199**
North West	1616	498	705	**2819**
Yorks and Humber	1354	420	530	**2304**
East Midlands	1383	350	445	**2178**
West Midlands	1422	520	580	**2522**
East	1556	436	515	**2507**
London	1778	715	1190	**3683**
South East	1925	636	965	**3526**
South West	1639	394	445	**2478**
Wales	779	332	240	**1351**
Scotland	1014	397	485	**1896**
Northern Ireland	1184	167	175	**1526**
UK	**16457**	**5032**	**6495**	**27984**

"Optimal National Allocation" by industry

	10-49	50-249	250+	Total
10-14	65	55	45	**165**
15-22	1140	489	720	**2349**
23-29	1434	621	790	**2845**
30-33	507	210	275	**992**
34-35	235	126	240	**601**
36-37	510	124	95	**729**
40-41	40	25	30	**95**
45	1714	332	350	**2396**
50-51	2334	631	595	**3560**
52	1445	211	540	**2196**
55	1965	364	350	**2679**
60-63	1004	241	380	**1625**
64.1	78	50	25	**153**
64.2	80	49	55	**184**
65-67	548	163	345	**1056**
70	365	111	140	**616**
71	176	64	60	**300**
72	390	150	145	**685**
73.1	61	51	60	**172**
73.2	40	0	0	**40**
74.2	356	93	110	**559**
74.3	60	45	15	**120**
rest of 74	1910	827	1130	**3867**
Total	**16457**	**5032**	**6495**	**27984**

Pros:
- Provides optimal national-level estimates.
- Consistent with CIS3 methodology.

Cons
- Regional imbalance, with larger regions having significantly larger samples than smaller regions and hence more precise estimates.

Scenario 3 – "Floor Allocation"

This scenario was carried out in the following stages:

- The sample for Wales, Scotland and Northern Ireland was as in scenario 1
- For the English regions, a "floor" sample size of 1,500 was initially set, in order to provide a minimum level of regional comparability
- The remainder of the sample (approximately 8,000 firms) was distributed across England using Neyman allocation

Table 8 – Sample design under scenario 3

"Floor Allocation" by region

	10-49	50-249	250+	Total
North East	1382	298	225	**1905**
North West	1359	446	705	**2510**
Yorks and Humber	1355	448	530	**2333**
East Midlands	1471	396	445	**2312**
West Midlands	1335	493	580	**2408**
East	1477	437	515	**2429**
London	1064	473	1190	**2727**
South East	1311	468	965	**2744**
South West	1589	405	445	**2439**
Wales	1271	504	240	**2015**
Scotland	1162	449	485	**2096**
Northern Ireland	1852	221	175	**2248**
UK	**16628**	**5038**	**6495**	**28161**

"Floor Allocation" by industry

	10-49	50-249	250+	Total
10-14	100	70	45	**215**
15-22	1,125	484	720	**2,329**
23-29	1,463	626	790	**2,879**
30-33	486	214	275	**975**
34-35	247	155	240	**642**
36-37	499	155	95	**749**
40-41	40	25	30	**95**
45	1,716	314	350	**2,380**
50-51	2,333	571	595	**3,499**
52	1,533	204	540	**2,277**
55	2,082	337	350	**2,769**
60-63	982	231	380	**1,593**
64.1	109	60	25	**194**
64.2	103	70	55	**228**
65-67	504	143	345	**992**
70	340	121	140	**601**
71	186	100	60	**346**
72	347	143	145	**635**
73.1	105	90	60	**255**
73.2	55	0	0	**55**
74.2	352	120	110	**582**
74.3	106	65	15	**186**
rest of 74	1,815	740	1,130	**3,685**
Total	**16,628**	**5,038**	**6,495**	**28,161**

Pros:

- Strikes a balance between regional and industrial/national accuracy.
- Data requirements in the devolved administrations maintained.

Cons:

- Small firms in SIC 50-51 (wholesale and automotive retail), 55 (hotels and restaurants) and "rest of 74" (other business activities excl. technical consultancy and testing) have exceptionally high samples, accounting for nearly a quarter of the overall sample.

Scenario 4 – "Hybrid Allocation"

This scenario was identical to scenario 3, but with the following additional stage:

- The sample of small firms in SIC 50-51, 55 and "rest of 74" was capped to 1,000. This freed up around 3,000 firms, which were distributed across the whole of the UK using Neyman allocation.

Table 9 – Sample design under scenario 4

"Hybrid Allocation" by region

	10-49	50-249	250+	Total
North East	1093	352	225	**1640**
North West	1370	528	705	**2598**
Yorks and Humber	1293	527	530	**2350**
East Midlands	1351	471	445	**2257**
West Midlands	1323	570	580	**2458**
East	1405	518	515	**2418**
London	1241	572	1190	**2988**
South East	1428	563	965	**2956**
South West	1456	474	445	**2360**
Wales	1012	558	240	**1790**
Scotland	1033	532	485	**2060**
Northern Ireland	1499	266	175	**1920**
UK	**15504**	**5931**	**6495**	**27930**

"Hybrid Allocation" by industry

	10-49	50-249	250+	Total
10-14	145	80	45	270
15-22	1329	572	720	2,621
23-29	1722	736	790	3,248
30-33	580	277	275	1,132
34-35	308	207	240	755
36-37	589	205	95	889
40-41	40	25	30	95
45	2024	384	350	2,758
50-51	1000	571	595	2,166
52	1785	265	540	2,590
55	1001	337	350	1,688
60-63	1156	294	380	1,830
64.1	169	65	25	259
64.2	158	85	55	298
65-67	601	211	345	1,157
70	417	176	140	733
71	246	150	60	456
72	433	194	145	772
73.1	155	110	60	325
73.2	60	0	0	60
74.2	425	177	110	712
74.3	161	70	15	246
rest of 74	1000	740	1,130	2,870
Total	**15,504**	**5,931**	**6,495**	**27,930**

Pros:

- As in scenario 3, but also ensures that small firms in SIC 50-51, 55 and "rest of 74" are not given unnecessarily large sample sizes.

Cons:

- Rather complicated!
- Reduces the sample size in Wales and Northern Ireland by approximately 10%.

Conclusion

It is recommended that the methodology described in scenario 4 be used for the UK CIS4 sample design. This will ensure that a decent level of accuracy can be achieved at the regional, industrial and national levels. Regional analysts will have substantial sets of data to enable local analysis, and the burden in the smaller regions and certain small firms will be reduced.

Recommendation: UK CIS4 sample to be based on design described in scenario 4

DTI-IG-TESE Jan 2005

Technical details of the UK Innovation Survey 2005 (CIS4)

Methodology

The UK Innovation Survey is funded by the Department of Trade and Industry (DTI). The survey was conducted on behalf of the DTI by the Office for National Statistics (ONS), with assistance from the Northern Ireland Department of Enterprise, Trade and Investment (DETINI).

The UK Innovation Survey is part of a wider Community Innovation Survey (CIS) covering European countries. The survey is based on a core questionnaire developed by the European Commission (EuroStat) and Member States. There have now been four innovation surveys (more detail can be found from the following link: www.cordis.lu/innovation-smes/scoreboard/home.htm).

The UK Innovation Survey 2005 (CIS4) sampled over 28 thousand UK enterprises. The survey was voluntary and conducted by means of a postal questionnaire.

Coverage and sampling

The survey covered enterprises with 10 or more employees in sections C-K of the Standard Industrial Classification (SIC) 2003. The 2005 survey included the following sectors: Sale, maintenance & repair of motor vehicles (SIC 50); Retail Trade (SIC 52); and Hotels & restaurants (SIC 55). These sectors were not included in the 2001 survey (CIS3). The 2005 survey sample was drawn from the ONS Inter-Departmental Business Register (IDBR) in December 2004.

Response and Weighting

The questionnaires from the initial survey were distributed on March 31 2005. Valid responses were received from 16,446 enterprises to give a response rate of 58 per cent.

Virtual Micro Data Laboratory Data Brief: Spring 2007[1]

Community Innovation Survey

Tomas Hellebrandt

The Community Innovation Survey (CIS) is a survey conducted every 4 years by EU member states to measure progress in the area of innovation. The CIS complements other indicators of innovativeness by providing a regular snapshot of innovation inputs and outputs and the constraints faced by businesses in their innovation efforts. This data brief provides an overview of the UK CIS and describes how the CIS data has been used for research in conjunction with other ONS business data sets held within the Virtual Microdata Laboratory. The brief concludes with an analysis of how various firm characteristics influence innovative activity.

1. Overview of the CIS

The CIS is based on a core questionnaire developed by the European Commission (EuroStat) and member States, to which the DTI adds questions for the purpose of the UK CIS. The survey structure has changed over time. In general the survey covers product, process and wider innovation including expenditure on different kinds of innovative activity, effects of innovation, sources of information and co-operation, barriers to innovation, protection methods for innovation, and public support for innovation.

Within the CIS**,** innovation is defined as ***major changes aimed at enhancing a firm's competitive position, performance, know-how, or capabilities for future enhancements***. These can be new or significantly improved goods, services or processes for making or providing them. Expenditure on innovative activities includes machinery and equipment, R&D, training goods and service design or marketing. The CIS is carried out at the level of the enterprise. As such, an enterprise may carry out one or more activities at one or more locations.

There have been four CIS surveys conducted to date, each covering a three-year period.

- CIS 1 covers the period 1991-1993. Due to the poor response rate (10%), this survey is regarded as being of poor quality and is not available within the VML.
- CIS 2 covers the period 1994-1996. In total 5,416 enterprises were surveyed, of which 2,339 responded to the survey achieving a response rate of 43%.
- CIS 3 covers the period 1998-2000 and was conducted in two waves. The first wave sampled 13,340 enterprises. Of the 19,625 enterprises to which the survey was sent, 8,172 responded achieving a response rate of 42%.

[1] For information on the CIS or other business data sets held within the VML, please email: vml@ons.gov.uk.

- CIS 4 covers the period 2002-2004. It is the largest of the innovations surveys conducted so far, sent to some 28,000 UK enterprises. Of those, 16,445 enterprises provided valid responses, representing a response rate of 58%.

The sample of enterprises is drawn from the ONS Inter-Departmental Business Register (IDBR) and is based upon those firms with more than 10 employees. The sample is designed to be statistically representative of the 12 regions of the UK, most industrial sectors and all sizes of firms. The responses to the survey are weighed back to the population using the inverse sampling proportion in each stratum. On average, each respondent in CIS4 represents 11 enterprises in the population.

The CIS is a voluntary postal survey. To boost response, enterprises are sent the survey, posted a reminder, posted a second reminder (with the survey again) and finally telephoned. There are a number of concerns about the data, primarily as the result of inconsistent responses provided by survey respondents. A number of marker variables were created in the cleaning process to identify problematic cases.

2. Linking CIS to ONS Business Data

The IDBR is the key sampling frame for business surveys within ONS. Enterprises appearing within ONS surveys are assigned a unique IDBR reference number which can facilitate linking of information on the same enterprise over time and between surveys. The reporting unit identifier in the CIS is given by ru_ref in CIS2 and ruref in CIS3 and CIS4. There are 789 reporting units which appear in CIS2 and CIS3, and 959 reporting units which appear in CIS3 and CIS4. One hundred and one reporting units appear in all three CIS surveys.

Linking of information on the same enterprise between surveys provides the opportunity to explore research questions that otherwise would not have been possible. The largest and most comprehensive ONS business survey is the Annual Business Inquiry. This survey includes information on turnover, costs, employment and investment. Due to the size and content of this survey, the ABI generally forms the spine against which most linking activity takes place. Responses to the CIS can therefore be linked to information collected on these organisations collected from the ABI.

Within the VML, information from the ABI is held in the Annual Respondents Database (ARD). To reduce compliance costs, the ABI is not a census of all businesses, with smaller reporting units being sampled. Within the ARD there are therefore two types of enterprise. Information collected directly from the survey returns of the ABI are held on the ‘selected files’ of the ARD. Information on those organisations included within the ABI survey universe but which are not included within the actual survey during a given year are held on the `non-selected’ files. By including information from the ‘non-selected’ ARD files, the coverage of the ARD is broadened considerably. However, the range of data items held on the non-selected files is more limited.

Table 1 shows the number of links that can be made between the CIS and the ARD, differentiating between links with selected and non-selected files. It is noted that whilst the ABI is an annual survey, the CIS covers a three year period. However, a number of questions in the survey refer only to the last of the three years covered by the survey. This suggests that the most appropriate link is likely to be to the ABI in that year. Table 1 therefore shows the number of links that can be expected when linking CIS2 to the 1996 ARD, CIS3 to the 2000 ARD and CIS4 to the 2004 ARD. It can be seen that for the CIS4, over 90% of enterprises can be linked to the ARD, with approximately 38% being able to be linked to the detailed information contained within the ‘selected’ files.

Table 1: CIS Survey Sample and links to the Annual respondents Database

	CIS Sample	Links to the Annual Respondents Database		
		Selected Files	Non-selected Files	Total Links
CIS2	2,342*	248	109	357
CIS3	8,172	3,472	4,010	7,482
CIS4	16,445	6,179	8,710	14,889

*Three replicated reference numbers have been removed from the original CIS sample

3. Previous Research Using the CIS

As the sponsor of the UK Community Innovation Survey, the DTI is one of the most active users of the survey for research purposes. A summary of some of the DTI-sponsored research projects using the CIS is shown below.

Innovative Business and the Science and Technology Base (Swann, 2002)
This analysis uses CIS3 to assess the role of the university (and other research institutions) as a source of information and cooperation for innovative businesses and the effect of such cooperation on business performance. It finds that companies are more likely to cooperate with universities when they are process innovators, but less likely when they are product innovators. The results suggest that universities play a relatively more important role through cooperation than as a source of information. Cooperation with the university is especially effective in achieving better process performance - i.e. greater production flexibility, reduced unit labour costs, and increased capacity.

Design and company performance: Evidence from the Community Innovation Survey (Cereda et. al, 2005)
This research undertakes an analysis of the relation between design inputs and other innovation and economic performance indicators. The authors find that around 9% of firms reported some spending on design and that design spending represented around 10% of all reported spending on innovative activities. They also find that design has a positive and statistically significant association with product innovation but not process innovation. They estimate a marginal return to design expenditure of about 17% which they state is likely to be an overestimate of the causal effect. Receiving government support raises design expenditure by about 3% of mean expenditure (for those firms undertaking expenditure).

Information Technology, Organisational Change and Productivity (Crespi, Criscuolo and Haskel, 2006)
This research uses the CIS3 and ARD to examine the relationships between productivity growth, IT investment and organisational change. Consistent with the small number of other micro studies the researchers find that (a) IT appears to have high returns in a growth accounting sense when organisational change is omitted; when organisation change is included the IT returns are greatly reduced, (b) IT and organisational change interact in their effect on productivity growth, (c) non-IT investment and organisational change do not interact in their effect on productivity growth.

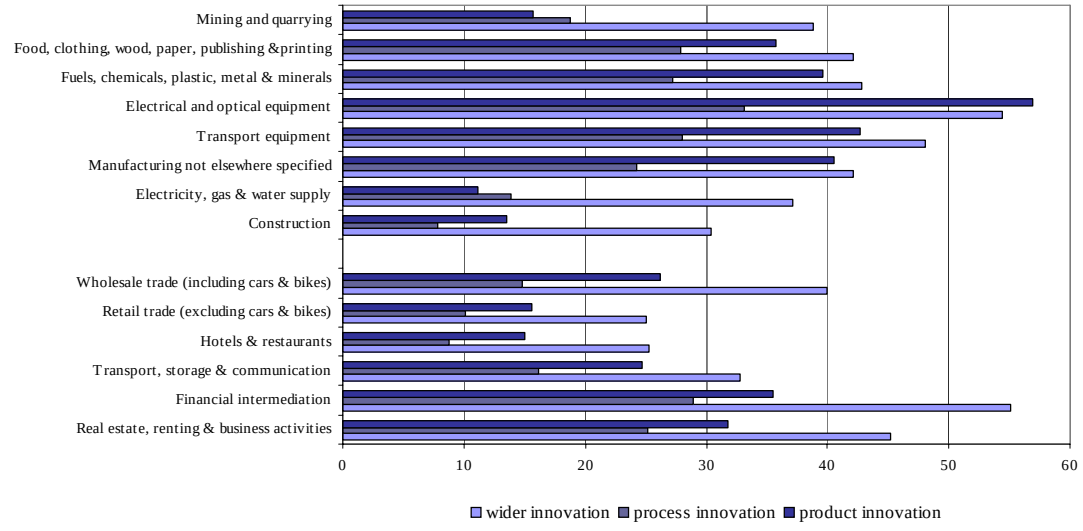
Productivity, Exporting and the Learning-by-Exporting Hypothesis (Crespi, Criscuolo and Haskel, 2006)
This research uses a matched CIS2-CIS3 panel to examine the proposition that exporting firms learn from their clients and this learning raises their productivity. The research finds that (a) firms who exported in the past are more likely to report that they learnt from buyers (relative to learning from other sources) and that (b) firms who had learned from buyers (more than they learnt from other sources) exhibited higher productivity growth, supporting the learning-by-exporting hypothesis.

Productivity Growth, Knowledge Flows and Spillovers (Crespi, Criscuolo, Haskel and Slaughter, 2007)
The linked CIS2-CIS3-ARD panel is used to explore the role of knowledge flows and TFP growth. Results suggest that the main sources of knowledge are competitors, suppliers, plants that belong to the same group and universities. They find a statistically significant association between TFP growth and above-firm-average information flows from other firms in the enterprise group, competitors and suppliers. The effects are economically significant as well, with such information flows "explaining" (in a growth accounting sense) about 50% of TFP growth. They conclude that the main "free" information flow spillover is from competitors and that multi-national presence may be a proximate source of this spillover.

4. Which firms innovate? Analysis of CIS4

The section describes the incidence of innovative[2] firms in the CIS4 according to a number of firm and market characteristics. Innovative activities are divided into three types, product process and wider innovation. Overall, 29% of firms in CIS 4 report product innovation, 20% report process innovation and 39% report wider innovation. Figure 1 shows that innovation activity is most prevalent in manufacturing. In construction, mining and quarrying, and electricity, gas and water supply innovative activity has a very different composition to that in other production sectors, dominated by wider innovation and process innovation outweighing product innovation in the latter two. Wider innovation also seems to be more prevalent, relative to product and process innovation in the distribution and service sectors.

Figure 1
Innovation activity by industry (%)



[2] Innovation is divided into product, process and wider innovation and the markers for these are available in the CIS4 data. Product innovation occurs when a firm introduces a new or significantly improved good or service. Process innovation occurs when a firm introduces new or significantly improved processes for producing or supplying goods or services and these processes are new to the enterprise. Wider innovation occurs when an enterprise makes major changes in business structure and practices, including corporate strategy, advanced management techniques, organisational structure and marketing.

Figure 2 shows that standalone enterprises are less likely to innovate than enterprises which are part of an enterprise group. This may partly reflect the fact that standalone enterprises are on average smaller (mean employment of 129 compared with 546 for enterprises that are part of a group), but it may also capture innovation spillovers between enterprises within an enterprise group. For all types of innovation, larger firms are more likely to innovate than smaller ones. Finally, innovation activity is found to be more prevalent among firms where at least one employee has a science or engineering degree, compared to other degrees and compared to enterprises where no employee has a degree[3].

Figure 2
Innovation activity by enterprise characteristics (%)

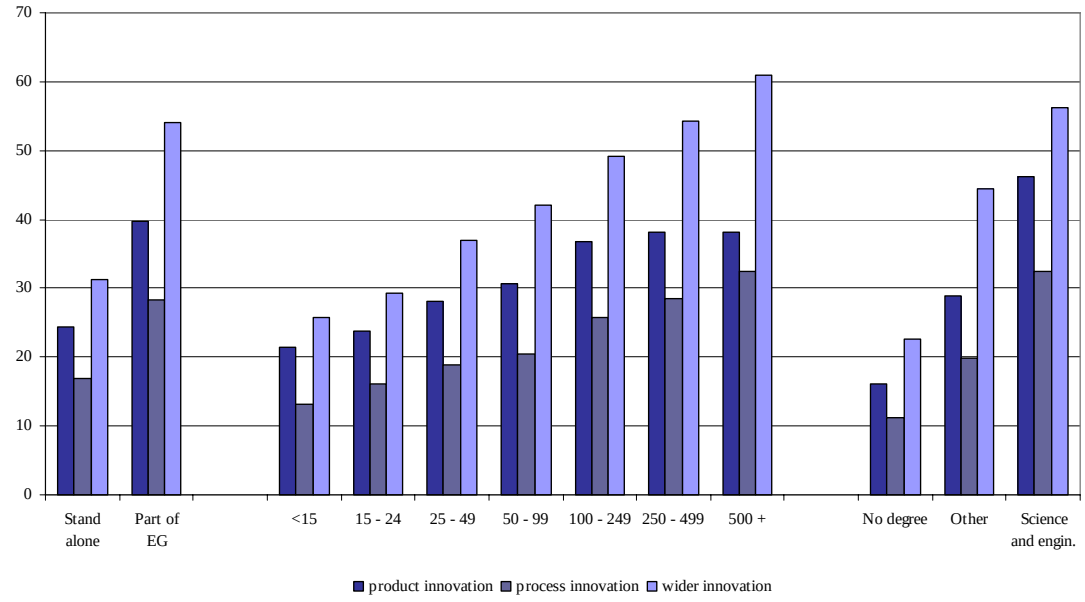


Figure 3 shows that firms engaging in export activities are far more likely to innovate in all areas, with a higher proportion of innovators among those firms who export beyond Europe. Innovation activity is most prevalent among firms whose main customers are other businesses, followed by those doing business with the public sector and finally consumers.

Comparing innovative activity between the three CIS surveys[4], Figure 4 suggests that innovative activity has increased in manufacturing between CIS3 and CIS4, whereas it has declined in construction, utilities and the distribution and service sectors. Electricity, gas and water supply is notable for the large fall in reported innovation over time. Figure 5 suggests that whilst innovative activity was relatively

[3] This variable is limited to a dummy designating whether or not the firm employs at least one employee with a degree. Respondents were asked to report the proportion of employees who have a degree but many appear to have reported the number of employees instead. The reported figures could therefore not be used to construct a more detailed measure of the education attainment of employees.

[4] Innovation-active firms in CIS3 and CIS4 are defined as those engaging in product, process or wider innovation, or innovation projects which have been abandoned or ongoing, or one or more innovative activities in the "Innovative activities and expenditures" part of the survey. The structure of the CIS2 questionnaire is quite different to the latter CIS surveys. In CIS2 innovation-active firms are defined as those that have undertaken service innovation, or innovation projects which have been unsuccessful, terminated, delayed or not yet completed (questions 3 and 13), or one or more organisational changes or new management techniques in question 4, or one or more innovative activities in question 7.

stable over the last 15 years amongst the smaller firms whereas it seems to have declined among the larger firms.

Figure 3
Innovation activity by market characteristics (%)

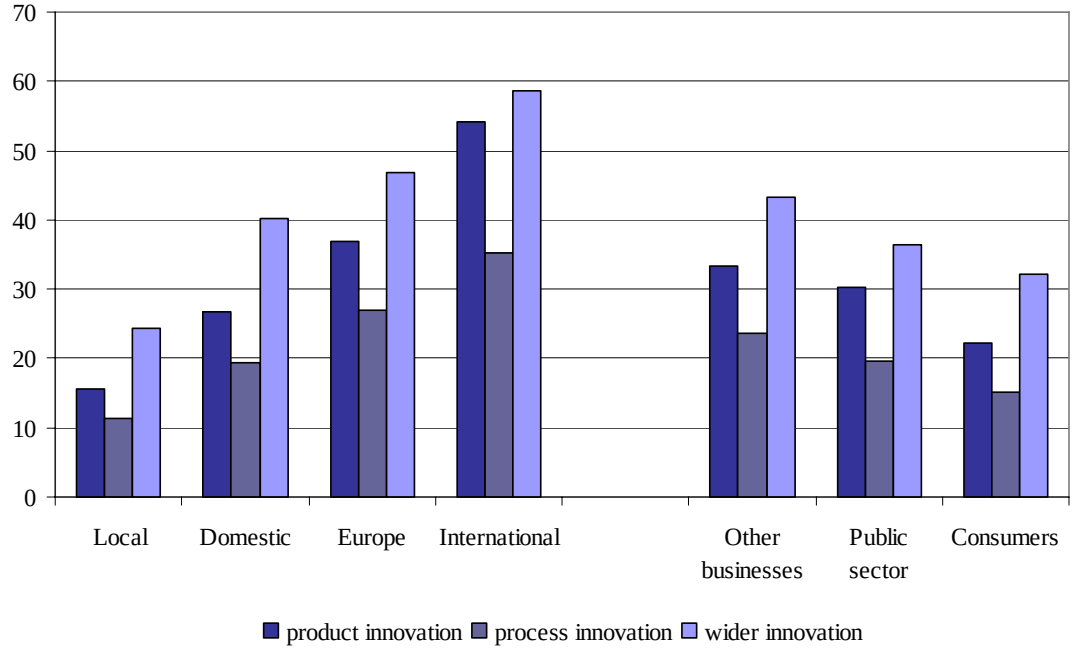


Figure 4
Changes in Innovation Activity: by industry (%)

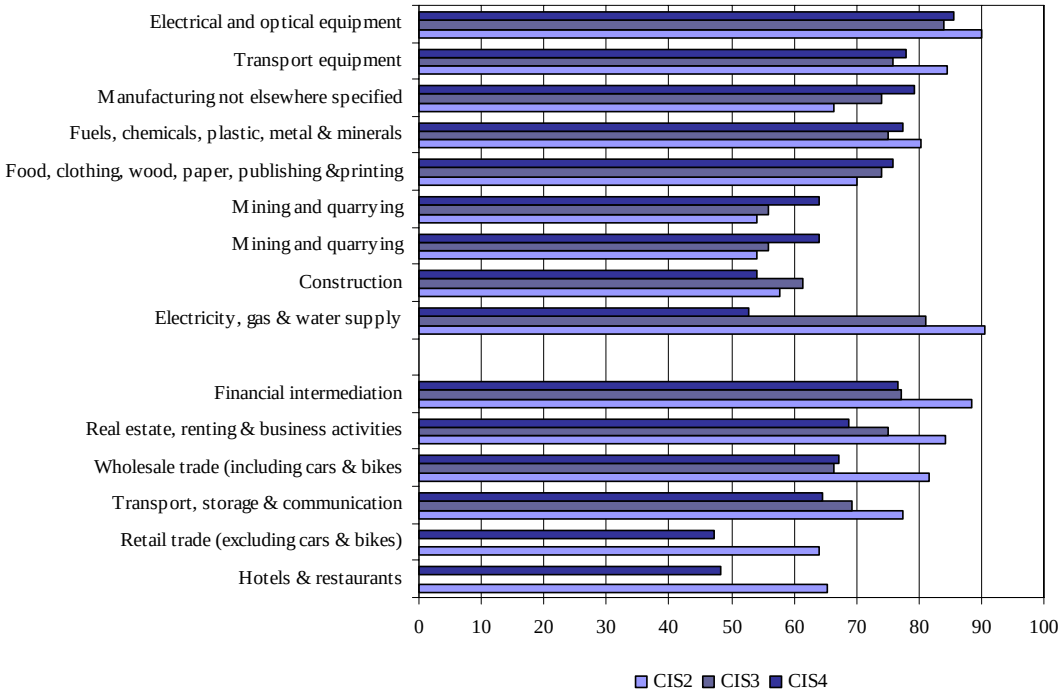
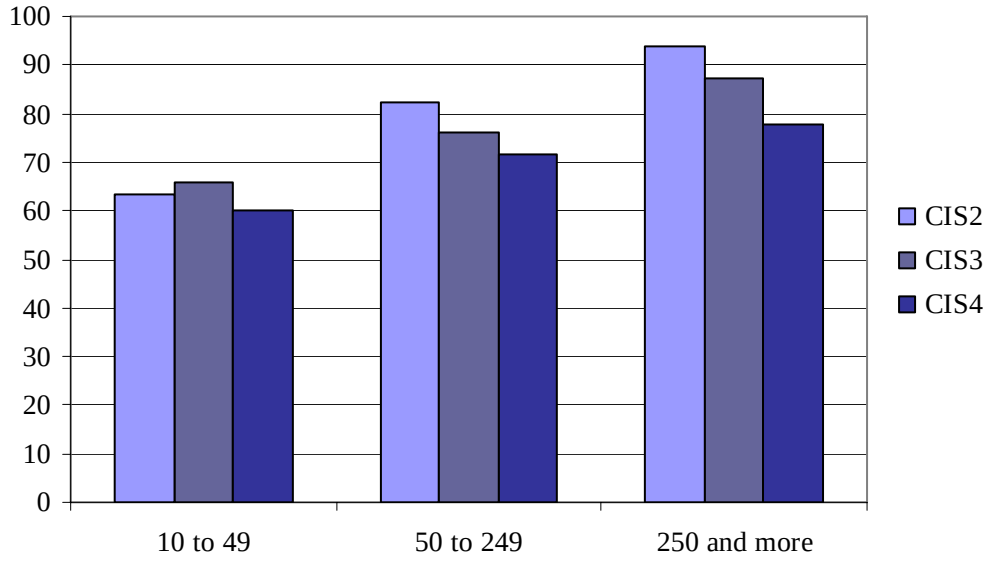


Figure 5
Changes in Innovation Activity: by size-band (%)



The above descriptive analysis of the CIS shows that innovation activity varies considerably across industry, size and other characteristics of the firm and market. However, such analysis does not identify the individual contribution of each of these characteristics on innovation, holding all other characteristics constant. For example, does being part of an enterprise group increase the likelihood of an enterprise engaging in innovative activity, or do these enterprises differ in other characteristics that may also make them more likely to innovate?

To do this, we employ logistic regression analysis to estimate the additional and independent effect of a range of firm and market characteristics on the probability that the firm will engage in innovative activity. We restrict the analysis to the most recent Community Innovation Survey. In addition to the information contained within the CIS, additional variables drawn from the 2004 Business Structure Database are merged onto the CIS data. The Business Structure Database is a version of the Inter Departmental Business Register held within the VML. Specifically, we use the BSD to explore how enterprise structure and company age effect innovative activity.

Explanatory variables are included to control for organisational structure, legal status, size, age, export market, main customers, industry and region. Selected results from the logistic regression are presented in Figures 6 and 7. It is noted that these results are derived from the same models presented in the Annex and are presented in separate charts purely for expositional convenience. For each of the variable sets, the results are expressed in terms of the percentage difference in the probability of engaging in innovative activities relative to a reference category. Innovation activities are once again divided into product, process and wider innovation. The coloured bars are used to indicate where a variable is estimated to be significantly different from the reference category at the 5% level.

Figure 6 confirms the earlier finding that larger firms are more likely to innovate, particularly in process and wider innovation. Amongst those organisations with less than 15 employees, the analysis distinguishes between those organisations where at least employee has a degree and those who do not. As noted above, we restrict this distinction to the smallest enterprises where it is more likely that the person filling out the survey will know if any of the employees has a degree. Among those enterprises with less than 15 employees, those enterprises with staff who have a science degree are approximately 200% more likely (or 3 times as likely) to engage in innovative

activity. The presence of staff with other degree subjects also increases the likelihood of innovative activity within small enterprises.

Figure 6
Probability of undertaking innovation activities: influence of firm size and education

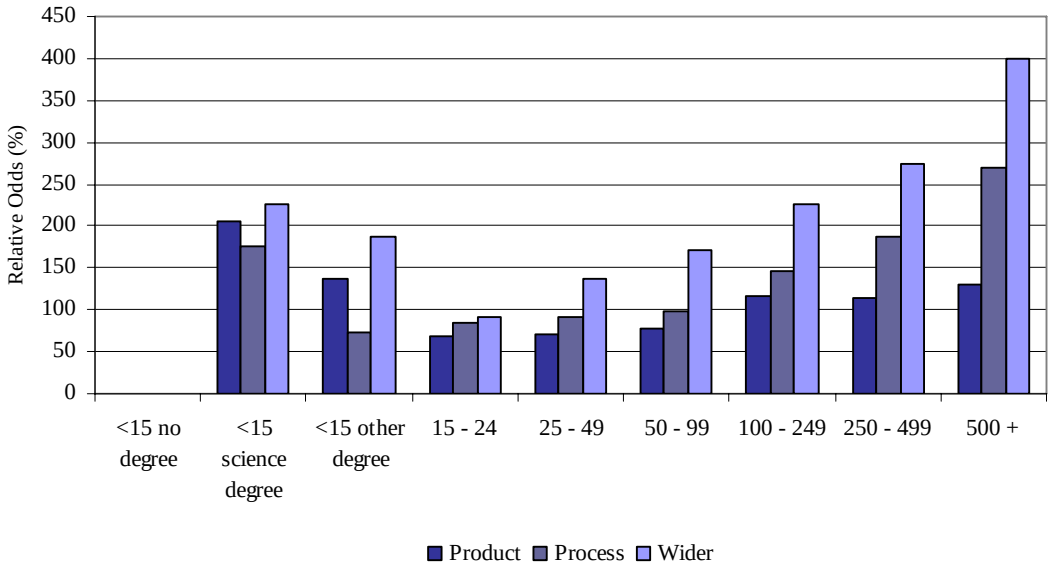


Figure 7
Probability of undertaking innovation activities: influence of organisational structure and age

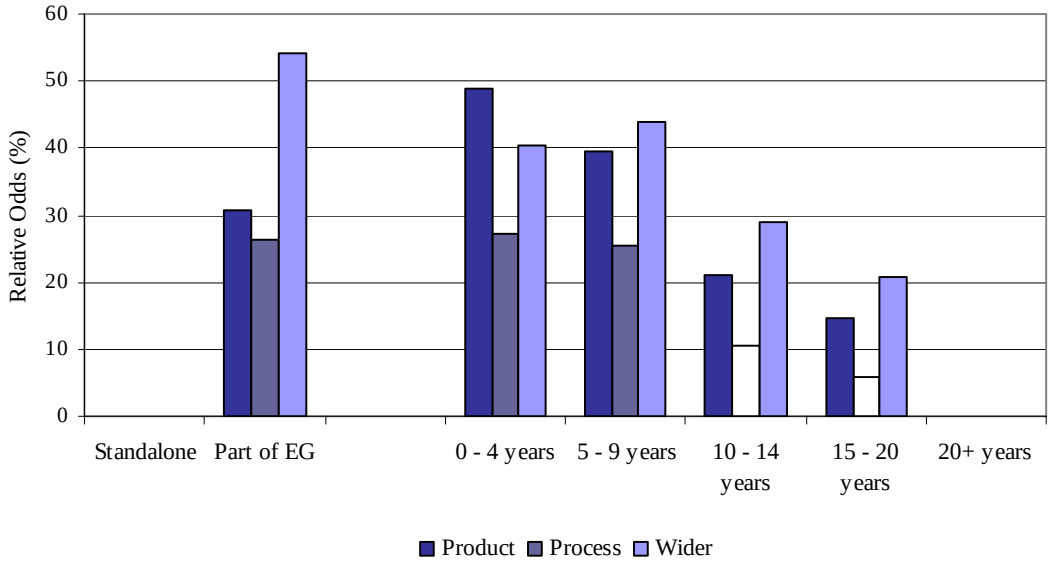


Figure 7 shows that being part of an enterprise group significantly increases the probability that a firm states that it engages in innovation, particularly wider innovation. Those enterprises who are part of an enterprise group are approximately 30% more likely to engage in product and process innovation, and 55% more likely to engage in wider innovation. Note that this effect is separate and additional to other characteristics of these enterprises. Finally, it is estimated that the youngest enterprises are most likely to state that they engage in innovative activity. Relative to those enterprises that have been established for 20 years or longer, those enterprises who have been established for 0-4 years are estimated to be 50% more likely to engage in product innovation, 27% more likely to engage in process innovation and 40% more likely to engage in wider innovation

Bibliography

Cereda et. al (2005), Design and company performance: Evidence from the Community Innovation Survey, Report for DTI

Crespi, Criscuolo and Haskel (2007), Information Technology, Organisational Change and Productivity, CEPR Working Paper

Crespi, Criscuolo and Haskel (2007), Productivity, Exporting and the Learning-by-Exporting Hypothesis, forthcoming Canadian Journal of Economics

Crespi, Criscuolo, Haskel and Slaughter (2007), Productivity Growth, Knowledge Flows and Spillovers, mimeo CeRiBA

Swann (2002), Innovative Business and the Science and Technology Base, Report for DTI

Please note: the variables included in the CIS4 & CIS2007 panel data are the same as those found in the CIS2007.

Also, a time variable is included in the panel data. This is labelled CIS_Version and equals 4 to indicate data from the CIS4, and 2007 to indicate data from the CIS2007. The data should therefore be sorted by ruref and CIS_Version.

Because not all of the questions in the CIS2007 are included in the CIS4, there will be missing values for some variables which were not included in the CIS4. Please note that approximately 7000 of the approx 15000 companies surveyed in each of the surveys have been matched to create the panel.

CLEANING THE CIS2

- Question 1b and 2b:
  - Replace prod_1 prod_2 prod_3 (procecss_1 process_2 and process3) with missing if prod_inn (proc_inn)=0.
- nov_inn: prod_inn is missing and nov_inn=1: these are service sector firms
- turn_nov: in 63 cases nov_inn=1 but turn_nov=.. 62 of these cases are in services, only 1 in manufacturing. PROBLEM: WHAT DO WE DO WITH THIS FIRM? It filled in the description of the innovation (dti3tx) but it did not fill in any of the turnover questions (turn_nov; turnnew; turnimp; turnung)
- ORGANISATIONAL CHANGE:
  - In 2 cases: everything is equal to zero, including intronon; which should be 1.
  - If intronon=1 then we should put all the others equal to 0 when missings
- FILTER QUESTIONS
  - Dti_inn1: if prod_inn or proc_inn or inun==1 then necessarily dti_inn1=1 therefore we replace dti_inn1=1 if dti_inn1==. & proc_inn=1 or productivity_inn=1 or inun=1
  - If dti_inn2 is not missing then dti_inn1 has to be zero; viceversa replace missing in dti_inn2 if dti_inn1==1
  - Lots of problems with consistency here: 3 firms say they do R&D continuously and 5 that they do it occasionally but they do not report R&D expenditure or R&D personnel.
  - In general, it should be the case that if firms answer no to 6b they do not answer the rest of the survey excluding question 14 onwards, but this is not always the case. THE QUESTION ARISE: what to do with these firms? Should we put their answers to missings?
- TURNOVER DISTRIBUTION:
  - Replace missings when we can calculate the percentage as residual from the other two categories.
- INNOVATION EXPENDITURES:
  - Replace 0 when it is missing if at least one of the other is not missing.
  - Create percentages of expenditures/turnover 1996 as reported in CIS2 and generate a marker (out_pc_`var') for those observations that have a percentage equal to or higher than 100%.
- R&D personnel
  - Create percentages of r&D personnel/employment 1996 as reported in CIS2 and generate a marker (out_pc_rdper) for the 1 observation that has a percentage equal to or higher than 100%. For now we mark observations/firms for which pc_rdper>100%. (for the analysis we are likely to drop them)
  - Create marker (weird_rdper)=1 if rdper>0 & intra_rd=0 or intra_rd=.; weird_rdper=2 if rdper=0 & intra_rd=1.
- R&D ACTIVITY
  - try to correct for inconsistencies: if you have not spent anything in R&D and do not have any R&D personnel how can you do R&D? We create a marker for these cases weird_rdcon=1 if intra_rd==0 & (extra_rd==0|extra_rd==.). Weird_rdcon=2 if rdcon<3 & intra_rd==0 & (extra_rd==0|extra_rd==.) & rdper>0 & rdper!=.

Ray explanation for this (Question: On Q,8 of CIS2 firms answer they are doing NO intramural R&D, but do employ personnel and do R&D continuously. Equally, there are firms who report they are doing intramural R&D, but they do NOT employ personnel and do NOT do R&D continuously.)Answer is that both groups are indeed engaged in intra-mural R&D. The first group have mostly just miscoded their answer to the Xinter question, though some may have used their R&D staff for other purposes in the accounting year in questions (must be a small %). The second are undertaking R&D in the survey year, but with technician staff, ather than dedicated R&D staff (hence they are not doing R&D continuously).

- Engineers in workforce
  - Use pc_qse as given in the survey that lies between 0 and 1. Checked with no_qse/emp_96 and it is fine.
- Factors influencing INNOVATIVE ACTIVITY
  - Sometines there are some "holes" in the data; i.e some anwers are ticked, others in the same observations are missings: we replace them with zero.
  - When they are all missings there is not much I can do (I leave them out)
  - Problem if both filter questions are "no" the firms should have not answered these questions, but they have. We create a dummy called filter for these observations and see what is the role they play in the analysis.

- SOURCES of INFORMATION for innovation projects (problems as before)
  - Sometines there are some "holes" in the data; i.e some anwers are ticked, others in the same observations are missings… we replace them with zero
  - When they are all missings there is not much I can do
  - Problem if both filter questions are "no" the firms should have not answered these questions, but they have. We create a dummy called filter for these observations and see what is the role they play in the analysis.

- SOURCES of TECHNOLGICAL KNOWLEDGE
  - Replace 0 when it is missing if at least one of the other is not missing.

- COOPERATION
  - In few cases cooperation=0 & partner of cooperation is not zero: we construct a marker for these cases : weird_coop _`var'=1. WE REPLACE COOPERATION=1 in these cases.

- FACTORS HAMPERING INNOVATION
  - Replace 0 when it is missing if at least one of the other is not missing.

- GOVERNMENT SUPPORT/GOVERNMENT PROGRAMS
  - PROBLEM: what do we do with missings in govsuppt and govprog? Do we replace them with 0s or do we leave them missings? I have checked and there are about 70 missings in manufacturing and only 32 are "justified" by the filter questions.
  - Replace 0 when it is missing if at least one of the other is not missing.

CLEANING FORREGRESSIONS

- Generate dummy for prod_1 prod_2 etc..=0 when prod_inn==0

POST-MERGING CLEANING:

- We check level of reporting unit using turnover and employment from ARD (and also location through region variable)

- We create a dummy for firms which report across groups

USE of NON-RESPONDENTS DATA

- Check characteristics of non-respondents
- Buid weights to reweight the data (basic idea: selectivity on observables).

Changes Applied to the Raw CIS3 Data

Authour: Brian Stockdale (DTI)

- When the data was originally received from ONS missing data was denoted with a –1. This was changed to SPSS's system missing as others in would ruin results.
- Where one part of a question had a response all the other unanswered parts were assumed to have an answer of well. This entailed changing non-response to nil response in some cases.
- Calculated percentage changes in employment, turnover to look for any unfeasible changes. Similar tests were carried out using turnover over employment and innovation expenditure over employment. Those showing blatantly unfeasible data were cross-checked against the actual images of the CIS forms and amended accordingly. The chief cause of this was unitary errors (£'s instead of £000's).
- Some forms displayed logical inconsistencies.
  - some enterprises claimed they received no financial assistance however also claimed that they took part in schemes which entailed financial assistance. This is a possible error with the questionnaire.
  - A number of firms also claimed that they engaged in innovation activity both continuously and occasionally.
- In the question on the percentage of employees holding degrees in 1) science and 2) other subjects, were the respondent has filled in one box but not the other we have entered a zero in the empty box.

Data Cleaning Applied to CIS3 Data 20/05/02

A number of respondents to the CIS3 questionnaire have entered their financial details incorrectly in that where we ask for units in thousands of pounds they have answered in pounds. It would be impossible to detect all of these but I have amended some of the blatantly obvious cases.

Within the dataset there are two variables, turn00 (turnover in 2000) and idbrturn (the enterprises turnover according to the inter-departmental business register in no specific year). Where turn00 is over 750 times greater than idbrturn I have divided all the financial data figures by a thousand. This has removed many of the unfeasible situations where a painting and decorating or plumbing firm was earning around £50m per man.

This method is by no means perfect and still leaves several firms that are very likely to have this error. However we cannot hope to track down all such errors and would like to avoid over-amending the raw data.

If, in the course of your research you find any *major* errors or ideas to improve the data quality that you feel we may want to know about please contact us.

For those of you who already had the data before the date above I enclose an annexe containing the reporting unit (ru) references of the cases amended and the variables which have been altered in those cases.

Cleaning/linking audit document

1. Basic information

Dataset: CIS3
Major version: 0
Minor version: 0

Document created: 17th June 2003 by Chiara Criscuolo
Supercedes: <nothing>

2. Associated documents

Changes applied to the raw CIS3 data.doc	Explains cleaning by DTI (official version)
CIS3 main questionnaire v2 nov version LABELS.pdf	Questionnaires and variable labels
CIS3_clean_1_0.do	Stata do file for cleaning
do_ard_2000.do	Stata do file for preparing ARD 2000
merge_cis3_ard_2000.do	Merge clean CIS3 with ARD 2000

3. General description

The Community Innovation Survey (CIS) is a voluntary postal survey carried out by ONS on behalf of the DTI. Eurostat proposes an initial questionnaire and the DTI adds questions. ONS randomly selects a stratified sample of firms with more than 10 employees, drawn from the Inter-Departmental Business Register (IDBR) by SIC92 2-digit class and 8 employment size bands. The survey covers both the production (manufacturing, mining, electricity, gas and water, construction) and the service sectors.

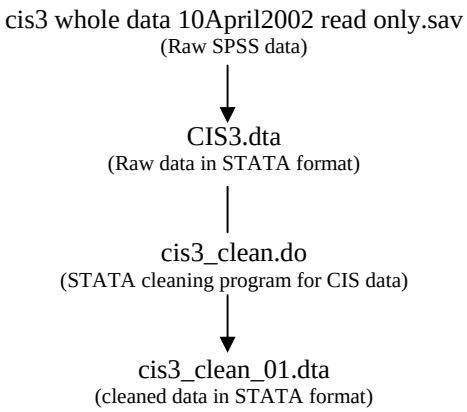
The Third Community Innovation Survey (CIS 3) was in the field twice. The first wave sampled 13,340 enterprises, the second top-up covered 6,285 to make the sample representative at the regional level. The CIS 3 covers the period 1998-2000. Of the total 19,625 enterprises to which the survey was sent, 8,172 responded (Table 1, row 1), achieving a response rate of 42%.

Table 1
CIS3 reporting unit profiles

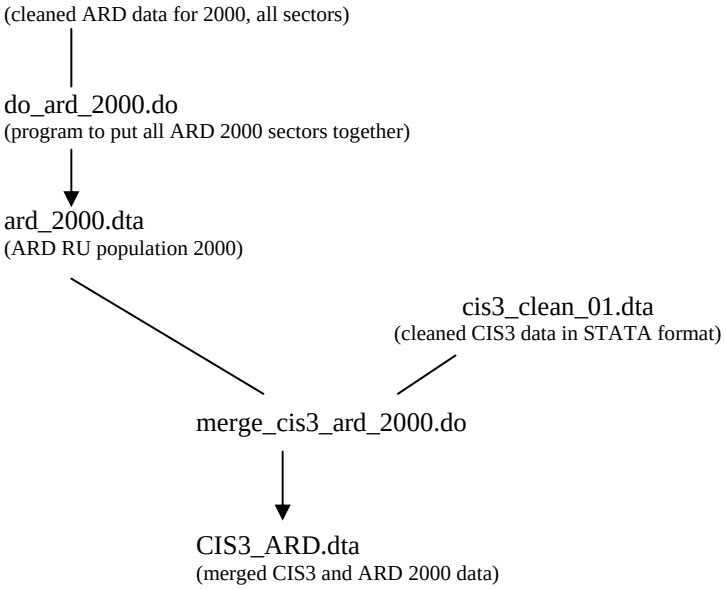
		CIS3
1	Number of Reporting Units	8,172
2	Number of Reporting Units in Services	3, 605
3	Number of Reporting Units in Production	4, 567

Source: Authors' calculations.

4. Program-datafile structure



dat2000xxx.dta, nul2000xxx.dta



The DTI sent the original data in SPSS format. We transferred it using Stat_Transfer into a new STATA dataset called *CIS3.dta*. We then cleaned this dataset using the file *cis3_clean.do*. The cleaned dataset is called *cis3_clean_01.dta*

At the same time we create the ARD "population" in 2000 using version 1.1 with the file called *do_ard_2000.do* and generate the dataset called *ard_2000.dta*

We match the cleaned CIS3 dataset (*cis3_clean_01.dta*) to the 2000 ARD dataset (*ard_2000.dta*) using the do file *merge_cis3_ard_2000.do*. The matched dataset is called *CIS3_ARD.dta*.

5. Detailed description of programs

The original SPSS file received by DTI has been transferred to STATA using STAT/Transfer. Care is needed so that the reporting unit and the enterprise identifiers are transferred as double variables.

5.1 CIS3_clean.do

Part 1: correcting problems identified by DTI

CIS3_clean.do checks the original CIS3 and implements some of the cleaning steps that the DTI documents report as being implemented but do not appear in the file we received. Namely:

- We replace numerical values (-1, and –2 for capex98) with missing values.
- The financial variables (**turn98 turn00 export98 export00 capex98 capex00 xinterm xextram xmachm xknowm xdesignm xtrainm xmarketm xtotalm**) have to be divided by 1,000 for the reporting units in the DTI document (we find another reporting unit for which this correction needs to be made and we add this to the list given to us by the DTI)
- We checked for duplicates. There are none.

Part 2: corrections from other consistency checks

- **Propsci/propoth**: in some cases one is missing but the other is not. In these cases, according to DTI guidelines, we replace missings with 0.
- If the variable **prodinov** (product innovation) is not missing, i.e. equals 0/1, and **prodnov** (new goods/services new to the market) is missing, the latter is replace with zero. Similarly for **procinov** (process innovation)
- We replace missings in the share of turnover (**sharenov**) due to products new to the market in 2000 with zero if **prodnov** equals 0.

- Wider innovation: we replace missings with zero following the DTI method: when an answer for another question is present (**orgstrat/orgmngt/orgorgan/orgmkt**)

part 3: generate useful variables

- generate RU reference as a string (for compatibility with other datasets)
- generate total production and expenditures as a % of turnover in 2000

part 4: generate inconsistency markers

CIS3_clean.do creates indicator variables for observations that present inconsistencies.

- Firstly, one of the main concerns is that reporting units are not reporting at the reporting unit level, i.e. they might be reporting at the local unit or at the enterprise level. This concern is caused by major differences between turnover and employment figures (both in 1998 and 2000) reported by the reporting units versus the figures from the IDBR register at the time the reporting unit was selected for the survey. We create an indicator variable for these observations. The dummy variable is called **checkturn** and it is equal to one every time the turnover reported by the reporting units is greater than 2*IDBR_turnover or less than 0.5* IDBR_turnover. Similarly a variable called **checkemp** was created for the employment variable.
- Innovation related expenditure in 2000:
  - We generated the sum of all the expenditure reported in the **survey (xinterm; xextram; xmachm; xknowm; xdesignm; xtrainm; xmarketm**) and compared it with the total expenditure in innovation as reported in the survey by the reporting units (xtotalm). In 3,359 cases they were equal, in 7 cases the sum we generated is less than the total reported by the firms, in 13 cases it is greater.
  - In 39 cases the total innovation expenditure is greater than total turnover in 2000. These observations have been marked with an indicator variable (**outpc**).
- There are some weird cases in the **rdpers** (persons involved in R&D activities within the enterprise in 2000). We flag those with the marker **outrdpers.** Outrdpers equals one if:
  - R&D personnel is greater than the number of total employees (9 cases)
  - R&D personnel is a positive number and employment in 2000 is missing (13 cases)
- Internal R&D. We flag inconsistencies between the various indicators of the R&D activities within the firm. These are **xinter; xinterm; xextra; xextram; rdpers; rdcont; rdoccas**. For the observations that present inconsistencies set **outrd=1** where:
  - R&D expenditure=0 even if no intramural R&D/R&D expenditure.
  - no R&D but R&D personnel
  - no internal or external R&D or personnel but R&D is being done continuously or occasionally
- Create marker for R&D done continuously or occasionally

Part 5: some further inconsistency checks done:

- **Prodnew/prodimp/produnc**: we check that the sum of the three is 100%, by creating a new variable prodtot, the sum of the three. Results were consistent, with almost all summing to 100, except a few cases where it was 99.
- How are the new products/processes developed? **prodwho** and **procwho**: in 15 cases prodinov equals 1 and prodwho is missing; in 9 cases procinov equals 1 and procwho equals missing. We have not flagged these observations but it is worth keeping this into account for the analysis.
- Innovation activities not completed or abandoned (**Aband/nyettime/nyetlate/noteven**): no case in which one part of the question had a response and another was unanswered. In 4,710 cases this information is missing.
- Factors hampering innovation: missing in 925 cases.
- **support**: there are various inconsistencies but perhaps, as noted by DTI is a problem with the survey itself and it is not possible to correct for those.

After the cleaning the dataset created is ready to be matched at the reporting unit level with any other dataset.

5.2 Matching the CIS3 with the ARD 2000

We match the cleaned CIS3 dataset to the 2000 ARD dataset at the reporting unit level using the do file *merge_cis3_ard_2000.do.* The matched dataset is called *CIS3_ARD.dta*. This contains matched observation and observations only present in CIS3.

Note that there are discrepancies in the sic92 classification as clear when comparing cis3_sic92 and sic92.

The dataset contains only few variables from the ARD, but it can be merged back to the ARD data to add information from the survey.

6. Reviewer's comments

Reviewed by: *Naveed Khawaja August 2003*

6.1 Information
CIS is quite complex, and users need to have a good look at the survey form before attempting to use the data.

6.2 Technical accuracy
Seems like all the obvious checks on the data have been done, plus other checks.

6.3 Semantic issues
None

6.4 Other comments
None

6.5 Other reviews

Reviewed by: *Felix Ritchie October 2003*

Found 2 errors in code; corrected while still at "approve" stage. No impact on published results (just indicator variables). Otherwise, concur with NK's comments and warnings about being wary of using the data.

Restructured CIS_Clean_1.0.do (original file was called CIS_clean.do) to make it readable. No effect on datasets.

7. User comments

7.1 Comment by <name> <date>

Cleaning/linking audit document

1. Basic information

Dataset: CIS3
Major version: 0
Minor version: 0

Document created: 17th June 2003
Supercedes: <nothing>

2. Associated documents

Changes applied to the raw CIS3 data	Explains cleaning by DTI (official version)
CIS3 main questionnaire v2 nov version LABELS.pdf	Questionnaires and variable labels
CIS3_clean.do	Stata do file for cleaning
Do_ard_2000.do	Stata do file for preparing ARD 2000
Merge_cis3_ard_2000.do	Merge clean CIS3 with ARD 2000

3. General description

The Community Innovation Survey (CIS) is a voluntary postal survey carried out by ONS on behalf of the DTI. Eurostat proposes an initial questionnaire and the DTI adds questions. ONS randomly selects a stratified sample of firms with more than 10 employees, drawn from the Inter-Departmental Business Register (IDBR) by SIC92 2-digit class and 8 employment size bands. The IDBR excludes agriculture, fishing and forestry, public administration and defence, education, health and social work. The survey covers both the production (manufacturing, mining, electricity, gas and water, construction) and the service sectors.

The Third Community Innovation Survey (CIS 3) was in the field twice. The first wave sampled 13,340 enterprises, the second top-up covered 6,285 to make the sample representative at the regional level. The CIS 3 covers the period 1998-2000. Of the total 19,625 enterprises to which the survey was sent, 8,172 responded (Table 1, row 1), achieving a response rate of 42%.
Of these 8,172 firms, 3,605 are in services and 4,567 in production.

Table 1

CIS3 reporting unit profiles

		CIS3
1	Number of Reporting Units	8,172
2	Number of Reporting Units in Services	3, 605
3	Number of Reporting Units in Production	4, 567

Source: Authors' calculations.

4. Program-datafile structure

The DTI sent the original data in SPSS format. We transferred it using Stat_Transfer into a new STATA dataset called **CIS3.dta**. We then cleaned this dataset using the file **cis3_clean.do**. The "cleaned dataset is called **cis3_clean_01.dta**

At the same time we create the ARD "population" in 2000 using version 1.1 with the file called **do_ard_2000.do** and generate the dataset called **ard_2000.dta**

We match the cleaned CIS3 dataset (**cis3_clean_01.dta**) to the 2000 ARD dataset (**ard_2000.dta**) using the do file **merge_cis3_ard_2000.do**. The matched dataset is called **CIS3_ARD.dta**.

5. Detailed description of programs

The original SPSS file received by DTI has been transferred to STATA using STAT/Transfer. Care is needed so that the reporting unit and the enterprise identifiers are transferred as double variables.

5.1 CIS3_clean

CIS3_clean checks the original CIS3 and implements some of the cleaning steps that the DTI documents reports as being implemented but do not appear in the file we received. Namely, the financial variables (turn98 turn00 export98 export00 capex98 capex00 xinterm xextram xmachm xknowm xdesignm xtrainm xmarketm xtotalm) have to be divided by 1,000 for the reporting units in the DTI document.

We find another reporting unit for which this correction needs to be made and we add this to the list given to us by the DTI.

We replace numerical values (-1, and –2 for capex98) with missing values.

We checked for duplicates. There are none.

This do file creates indicator variables for observations that present inconsistencies.

Firstly, one of the main concerns is that reporting units are not reporting at the reporting unit level, i.e. they might be reporting at the local unit or at the enterprise level. This concern is caused by major differences between turnover and employment figures (both in 1998 and 2000) reported by the reporting units versus the figures from the IDBR register at the time the reporting unit was selected for the survey. We create an indicator variable for these observations. The dummy variable is called **checkturn98 (checkturn00)** and it is equal to one every time the turnover reported by the reporting units is greater than 2*IDBR_turnover and less than -0.5* IDBR_turnover. Similarly a variable called **checkemp98 (checkemp00)** was created for the employment variable.

Secondly, we looked at presumably incredible values for turnover per employee figures. We generated a dummy variable **checkte98 (checkte00)** equal to one if the difference between the turnover per employee as reported by the reporting unit is greater than 2 times or less than half the turnover per employee figure calculated using the IDBR values.

We then went on to check other variables in the dataset. We describe the details for each of them:

- Propsci/propoth: in some cases one is missing but the other is not. In these cases, according to DTI guidelines, we replace missings with 0.
- Prodnew/prodimp/produnc: we check that the sum of the three is 100%, by creating a new variable prodtot, the sum of the three.
- If the variable prodinov (product innovation) is not missing, i.e. equals 0/1, and prodnov (new goods/services new to the market) is missing, the latter is replace with zero. Similarly for process innovation
- We replace missings in the share of turnover due to products new to the market in 2000 with zero if prodnov equals 0.
- How are the new prodducts/processes developed? Prodwho and procwho: in 15 cases prodinov equals 1 and prodwho is missing; in 9 cases procinov equals 1 and procwho equals missing. We have not flagged these observations but it is worth keeping this into account for the analysis.
- Innovation activities not completed or abandoned (Aband/nyettime/nyetlate/noteven): no case in which one part of the question had a response and another was unanswered. In 4,710 cases this information is missing.
- Factors hampering innovation: missing in 925 cases.
- Innovation related expenditure in 2000:
-
    - We generated the sum of all the expenditure reported in the survey (xinterm; xextram; xmachm; xknowm; xdesignm; xtrainm; xmarketm) and compared it with the total expenditure in innovation as reported in the survey by the reporting units (xtotalm). In 3,359 cases they were equal, in 7 cases the sum we generated is less than the total reported by the firms, in 13 cases it is greater.
    - In 39 cases the total innovation expenditure is greater than total turnover in 2000. These observations have been marked with an indicator variable (**outpc**).
    - There are some weird cases in the rdpers (persons involved in R&D activities within the enterprise in 2000). We flag those with the marker **outrdpers.** Outrdpers equals one if:

- in 5 cases R&D personnel is greater than the number of total employees and in 4 cases R&D personnel is a positive number and employment in 2000 is equal to 0. (in 7 cases they amay be explained by the firm having incurred a structural change Finally in 13 cases R&D personnel is a positive number and employment in 2000 is missing.
- Internal R&D. We flag inconsistencies between the various indicators of the R&D activities within the firm. These are xinter; xinterm; xextra; xextram; rdpers; rdcont; rdoccas. For the observations that present inconsistencies we construct an indicator variable called **outrd**. Outrd takes value one in the following cases:

- R&D expenditure=0 even if in Question 9.1 no intramural R&D and/or no intramural R&D expenditure.
- Question 10.2: how did your enterprise engage in R&D during the three year period? In 8 cases the variables RDCONT and RDOCCAS are equal to 1 even firms does not do any internal R&D according to question 9.1. In 7 cases there are no R&D personnel no expenditure in R&D but R&D is done continuously; in 1 case the R&D is done extramurally

- SUPPORT: there are various inconsistencies but perhaps, as noted by DTI is a problem with the survey itself and it is not possible to correct for those.
-
- Wider innovation: we replace missings with zero following the DTI method: when an answer for another question is present

After the cleaning the dataset created is ready to be matched at the reporting unit level with any other dataset.

5.2 Matching the CIS3 with the ARD 2000
We match the cleaned CIS3 dataset to the 2000 ARD dataset at the reporting unit level using the do file **merge_cis3_ard_2000.do**. The matched dataset is called **CIS3_ARD.dta**. This contains matched observation and observations only present in CIS3.

Note that there are discrepancies in the sic92 classfication as clear when comparing cis3_sic92 and sic92.

The dataset contains only few variables from the ARD, but it can be merged back to the ARD data to add information from the survey.

6. Reviewer's comments

Reviewed by: <name> <date>

6.1 Information

6.2 Technical accuracy

6.3 Semantic issues

6.4 Other comments

7. User comments

7.1 Comment by <name> <date>